Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Incremental prognostic value of early arterial blood gases at ICU reassessment in adults with hypercapnia: a development and external validation study

Jitao Xu
Shandong University of Traditional Chinese Medicine
Xitong Liang
Shandong University of Traditional Chinese Medicine
Peng Li
Shandong University of Traditional Chinese Medicine
Xinrui Cai
Shandong First Medical University
Weibin Qian
doctorqiantcm@126.com

Affiliated Hospital of Shandong University of Traditional Chinese Medicine

Research Article

Keywords: hypercapnia, arterial blood gas, intensive care, mortality prediction, external validation, explainable machine learning

Posted Date: November 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11318566/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Arterial blood gases are routinely repeated to monitor therapeutic response in critically ill patients with hypercapnia. However, whether initial acid-base history retains incremental prognostic value once contemporary clinical status and reassessment blood gases are known remains unclear. We evaluated whether early pH and $PaCO_2$ improve in-hospital mortality prediction beyond current physiological information at predefined ICU reassessment landmarks.

Methods

We developed models in MIMIC-IV 3.1 and evaluated frozen predictions in eICU 2.0. Adults had $PaCO_2$ >45 mmHg in their first eligible paired arterial blood gas within 6 hours of ICU admission. The primary 24-hour landmark required survival, continued ICU stay and an 18–24-hour repeat. Five algorithm families were compared using patient-grouped nested cross-validation. The current-information model (L) and history-augmented model (H) received 11 clinical inputs and the latest pH/$PaCO_2$; H also received the first pH/$PaCO_2$. The primary difference was Brier(L)−Brier(H), with hospital–patient product-weight confidence intervals. Secondary analyses examined 12-hour reassessment and a 24-hour fallback workflow.

Results

Development included 2662 admissions with 561 deaths; external evaluation included 2506 admissions with 577 deaths across 154 hospitals. External H/L Brier scores were 0.1476/0.1478 and AUCs 0.7545/0.7536. The paired Brier difference was 0.27×10^{-3} (95% CI -1.10×10^{-3} to 1.69×10^{-3}). Calibration slopes were 0.880/0.882. The 12-hour difference was 0.75×10^{-3} (95% CI -1.78×10^{-3} to 3.51×10^{-3}). The external workflow used H in 23.06% of 10865 admissions and yielded a Brier difference of -0.12×10^{-3} (95% CI -2.17×10^{-3} to 1.90×10^{-3}). SHAP attributed greater average importance to current pH and clinical status than to early gas values. Browser calculations were verified against the Python implementation.

Conclusions

Early blood gas history provided an uncertain predictive increment after contemporary reassessment. Repeat measurement coverage and external calibration constrain application of these landmark models. The verified browser implementation supports research inspection of the frozen predictions.

Background

Arterial blood gases describe the respiratory and acid-base state of critically ill patients. Their prognostic interpretation depends on the joint values of pH and arterial carbon dioxide tension ($PaCO_2$). Clinical guidelines use respiratory acidosis to guide noninvasive ventilation in COPD exacerbations and recommend reassessment of physiological response during treatment [1, 2]. In a large cohort of mechanically ventilated patients, Tiruvoipati and colleagues found different mortality associations for compensated hypercapnia and hypercapnic acidosis [3]. Studies in moderate or severe acute respiratory distress syndrome have also associated severe or sustained hypercapnia with ICU mortality [4, 5]. These observations motivate joint assessment of pH and $PaCO_2$, while their clinical meaning also depends on the patient's disease, metabolic state and ventilatory management.

Blood gas measurements already contribute to respiratory prognostic models. The COPD-specific HACOR score combines acidosis with heart rate, consciousness, oxygenation and respiratory rate to predict noninvasive ventilation failure, particularly after 1–2 hours of treatment [6]. The NIVO score predicts in-hospital mortality in exacerbations of COPD requiring assisted ventilation, using pH together with pre-admission functional limitation, clinical findings and the timing of acidaemia [7]. In public ICU databases, Lu and colleagues developed a Charlson comorbidity index-based machine learning model for reported 28-day mortality in acute hypercapnic respiratory failure [8]. These models establish the relevance of blood gases or broader clinical burden for their specified tasks. They also demonstrate why prediction time, patient selection and outcome definition must accompany comparisons between models.

Repeated measurements raise a conditional prediction question about the information retained in earlier blood gases once the patient's current state is known. Landmark prediction specifies the time at which prognosis is updated and the population still at risk [9]. Zong and colleagues evaluated longitudinal blood gas information at day 3, 5 and 7 landmarks in sepsis using MIMIC-IV and eICU [10]. Liu and colleagues developed a two-hour reassessment model for noninvasive ventilation failure in hypercapnic COPD, including a comparison with reassessment information alone [11]. At an ICU reassessment, adding earlier measurements may capture the course preceding the current state. Its incremental value can be estimated by giving both models the same latest blood gas and clinical information. Association with mortality alone is insufficient to establish added predictive ability [12]. With two measurements, the first and latest values contain the information needed to calculate their change; a longer series is required to describe a detailed trajectory.

This comparison also depends on how measurements enter the record. Repeat blood gas sampling follows clinical decisions, and availability can reflect illness severity or local workflows. Agniel and colleagues demonstrated that healthcare processes themselves can contribute to the predictive information in electronic health records [13]. External evaluation therefore needs a defined risk set, harmonised predictors and assessment of absolute-risk calibration as well as discrimination [14]. TRIPOD + AI and PROBAST + AI provide guidance for transparent reporting and appraisal across regression and machine learning methods [15, 16]. A paired comparison of probability errors can assess

the added information directly; the Brier score evaluates probabilistic accuracy alongside complementary measures of discrimination and calibration [17, 18].

We used MIMIC-IV 3.1 for model development and eICU for independent external evaluation [19, 20, 21]. Among adults with an early paired arterial blood gas showing $PaCO_2$ >45 mmHg, we compared models using the latest pH/$PaCO_2$ and shared clinical inputs with models that also included the first pH/$PaCO_2$. The primary assessment concerned subsequent in-hospital death at the 24-hour ICU landmark. Prespecified analyses examined a separate 12-hour risk set and a 24-hour workflow that retained a first-measurement model when a repeat was unavailable. We also examined measurement coverage across hospitals, explained the frozen predictions and tested their implementation in a browser research calculator.

Methods

Study design and population

We conducted a retrospective prediction study using MIMIC-IV version 3.1 (2008–2022) for development [19, 20] and the eICU Collaborative Research Database version 2.0 (2014–2015) for external validation [21, 22]. Both resources are hosted by PhysioNet [23]. The prespecified question was whether early arterial blood gas history added prognostic information beyond the latest blood gas and recorded clinical status. Reporting was structured around TRIPOD + AI [15]. Source archives were processed without modification, and patient-linked intermediate data remained in the local analysis workspace.

Eligible patients were adults aged at least 18 years in the first ICU stay of each hospitalization, with $PaCO_2$ >45 mmHg in the earliest qualifying paired arterial pH/$PaCO_2$ episode within 0–6 hours of ICU admission. An unresolved earliest candidate was excluded without selecting a later episode. Repeated hospitalizations were retained and grouped by patient during development. Cohort inclusion used the blood gas definition and did not require a respiratory-failure diagnosis.

The primary landmark was 24 hours after ICU admission [9]. Patients had to remain in the ICU beyond that time and meet survival and outcome-consistency rules based on available records. The latest candidate paired gas in the inclusive 18–24-hour window supplied the current blood gas. Conflicting latest candidates were treated as unavailable without returning to an older candidate. The outcome was subsequent death during the same hospitalization. MIMIC death and discharge timestamps were checked against the landmark; eICU unit/hospital discharge offsets and discharge status supplied timing proxies. Unknown outcomes were excluded from performance evaluation and counted separately.

Measurement definitions and predictor availability

MIMIC pH and $PaCO_2$ were paired by specimen identifier and chart time with an arterial specimen designation. In eICU, blood gas laboratory category and identical result offsets defined a pair; specimen

identifiers were unavailable. Duplicate identical values were retained as one episode, and disagreement in values or arterial classification made the episode unresolved. Both early and repeat analytes required plausible record timestamps between their event and the landmark. MIMIC used chart time and store time; eICU used result and revision offsets. Revision offsets were a conservative availability proxy, and the exported tables did not reconstruct every earlier result version.

Both models used 11 shared clinical inputs: age, recorded sex, heart rate, respiratory rate, mean arterial pressure, peripheral oxygen saturation, lactate, creatinine, the time since each laboratory measurement, and a comparable Glasgow Coma Scale (GCS) total. Age was capped at 90 years in both databases. Unknown sex remained missing. Vital signs used the latest eligible value in the preceding 6 hours. Lactate and creatinine used the latest eligible result since ICU admission, with elapsed measurement time entered separately. Units, source whitelists and broad physiological bounds were audited before modeling. An invalid latest vital sign or an unresolved same-time conflict was retained as missing without older-value substitution. GCS required a valid total or complete, consistent eye/verbal/motor components; unassessable components, including an intubated verbal response, remained missing. Records with entry times required event ≤ entry ≤ landmark. eICU periodic and aperiodic vital signs had event-time checks without independent entry timestamps.

Current respiratory support was excluded from the shared primary input set before performance analysis because the initial cross-database mapping could not establish comparable states. Further source auditing informed a separate support sensitivity analysis. COPD and respiratory-failure codes from the hospitalization were used only for retrospective descriptive subgroups.

Model development and selection

The current-information model, L, combined the shared clinical inputs with the latest pH and $PaCO_2$. The history model, H, added the early pH and $PaCO_2$. An independently selected current-information model, L_best, provided an additional comparator. All comparisons used identical admissions and patient-grouped partitions.

We compared penalized spline logistic regression, random forest [24], histogram gradient boosting (HistGB), based on the gradient-boosting framework [25], XGBoost [26] and a radial-basis-function support vector machine [27]. Each family had three prespecified configurations. Operational definitions and configuration details are provided in Additional file 1. Clinical variables were median-imputed within training folds, with fixed missingness indicators for every clinical input; an entirely missing training column used zero with its indicator. Blood gases required complete qualifying pairs. Scaling was fitted within the logistic and support-vector pipelines. Logistic models used restricted cubic spline blood gas terms with training-derived knots at the 10th, 50th and 90th percentiles and a centered, scaled latest-pH–$PaCO_2$ product. This gave L flexibility for nonlinear current blood gas associations. Every candidate underwent sigmoid probability calibration [28, 29] using three patient-grouped folds within its training

set, followed by a full-training-set base-estimator fit. We used neither synthetic oversampling nor class weighting.

Development used five outer folds and three inner folds with stratified patient-grouped allocation. Within each outer training set, minimum inner Brier score selected the H family and configuration; L was tuned independently within that family, and L_best was selected independently across families. Outer predictions therefore evaluated the complete selection procedure, with model selection repeated within each outer training set [30]. Per-family outer predictions were also summarized descriptively. Final configurations were selected using five grouped folds in the complete MIMIC cohort, fitted with the same internal calibration recipe, and frozen before eICU evaluation. External outcomes did not inform selection, tuning or recalibration.

Performance and prespecified secondary analyses

The primary estimand was the admission-level paired difference ΔBrier = Brier(L)−Brier(H), with positive values favoring H. We also reported absolute Brier score, log-loss, area under the receiver-operating-characteristic curve (AUC), calibration-in-the-large, jointly estimated calibration intercept and slope, and prediction-quantile calibration plots [18, 31]. External calibration regressions were diagnostic and left predictions unchanged. External intervals used independent Poisson(1) weights for hospitals and patients, multiplied for each admission and normalized within each draw. Percentile 95% intervals used 2000 draws for Brier/log-loss and 500 for AUC/calibration. These intervals condition on frozen predictions and represent sampling variability in the validation hospitals and patients; they exclude model redevelopment uncertainty. Hospital coverage and performance were described without fitting separate calibration slopes in sparse sites.

The 12-hour analysis used a separately defined landmark risk set and latest qualifying paired blood gases at 8–12 hours. Its H/L configurations and calibration were developed separately in MIMIC within the HistGB family selected by the primary analysis. A prespecified wider-window sensitivity applied the frozen 12-hour models to the latest gas in (6,12] hours. The 24-hour availability-based workflow used a penalized spline logistic model, B_all, trained in the broader landmark population with clinical inputs and the early gas. It used H24 when a reliable repeat was available and B_all otherwise. Workflow development formed common patient-grouped outer folds before fitting the all-patient fallback and paired-patient update branches. Secondary internal estimates condition on the primary family choice made in overlapping MIMIC data. The support sensitivity added the same four mutually exclusive support indicators to H and L, with fold-local tuning and calibration. Acidic early pH, COPD codes and respiratory-failure codes informed descriptive subgroup analyses.

Model explanation and browser implementation

Permutation SHAP explained the logit of the complete calibrated H probability pipeline, including missing-data handling [32]. We randomly selected 32 MIMIC background admissions and 128 eICU explanation admissions. Checks included additivity, a disjoint background of 32 admissions on the same 64 cases, 1000 explanation-sample resampling draws, and a prespecified blood gas grouping sensitivity. These assessed explanation computation and sample/background sensitivity. Conditional joint-gas plots used synthetic clinical profiles and a development-derived gas-support mask based on tenth-neighbor distances.

We exported the frozen H24/L24, H12/L12 and B_all24 pipelines to JavaScript. A publicly accessible interface loads a local model JSON through the browser File API, checks its SHA-256, structure and version, and computes in memory. Hosted assets contain generic inference code and an aggregate SHAP figure; model parameters are supplied separately under source-database sharing conditions [33]. Engineering evaluation included JavaScript–Python numerical comparisons in Node.js, automated application checks, three synthetic routes in an actual browser, and one mobile-width check. The calculator supports paired comparisons at both landmarks and the 24-hour fallback; the 12-hour route requires a qualifying repeat.

Blood gas and pressure variables retain the source clinical unit mmHg (1 mmHg = 0.1333224 kPa); creatinine retains mg/dL (1 mg/dL = 88.4 μmol/L). These conversions provide the corresponding SI units without changing the fitted model input scale.

Python 3.10.4 and related computational packages were used for statistical modeling and analysis. Generative AI tools (OpenAI) were used solely for code script debugging, figure formatting, and language editing of the authors' original manuscript draft. All research hypotheses, clinical definitions, analytical workflows, data verification, and manuscript conclusions were independently conceived, verified, and approved by the authors; no large language model processed individual patient records or classified clinical outcomes.

Results

Cohort composition and measurement coverage

The primary MIMIC cohort comprised 2662 admissions from 2609 patients, with 561 subsequent in-hospital deaths (21.1%). External validation comprised 2506 admissions from 2456 patients across 154 hospitals, with 577 deaths (23.0%); 22 otherwise qualifying paired admissions had unknown outcomes. Cohort selection is shown in Fig. 1, and characteristics and variable-specific missingness are reported in Table 1.

Table 1
Primary paired cohort characteristics

Characteristic	MIMIC-IV n = 2662	Missing	eICU n = 2506	Missing
Age (years; capped at 90)	64.0 (54.0–74.0)	0.0%	65.0 (54.0–74.0)	0.0%
Male sex	1744/2662 (65.5%)	0.0%	1417/2504 (56.6%)	0.1%
Heart rate (bpm)	86.0 (74.0–98.0)	0.6%	87.0 (74.0–101.0)	0.7%
Respiratory rate (breaths/min)	20.5 (17.0–25.0)	1.2%	20.0 (16.0–25.0)	9.7%
Mean arterial pressure (mmHg)	74.0 (68.0–84.0)	6.8%	77.0 (68.0–88.0)	0.8%
SpO_2 (%)	97.0 (95.0–99.0)	0.9%	97.0 (95.0–99.0)	1.5%
Lactate (mmol/L)	1.6 (1.1–2.3)	3.8%	1.7 (1.1–3.0)	50.8%
Creatinine (mg/dL)	1.1 (0.8–1.8)	0.3%	1.2 (0.8–2.0)	7.9%
Lactate measurement age (h)	4.3 (2.0–10.1)	3.8%	9.2 (3.5–19.8)	50.8%
Creatinine measurement age (h)	6.7 (3.9–10.1)	0.3%	7.2 (4.0–11.8)	7.9%
Comparable GCS total	15.0 (14.0–15.0)	79.9%	11.0 (8.0–15.0)	45.3%
First pH	7.28 (7.21–7.34)	0.0%	7.28 (7.21–7.33)	0.0%
First $PaCO_2$ (mmHg)	52.0 (48.0–60.0)	0.0%	55.2 (49.0–68.0)	0.0%
Current pH	7.37 (7.32–7.42)	0.0%	7.37 (7.31–7.42)	0.0%
Current $PaCO_2$ (mmHg)	42.0 (37.0–49.0)	0.0%	45.0 (38.0–53.9)	0.0%
pH change (current − first)	0.09 (0.03–0.16)	0.0%	0.09 (0.03–0.16)	0.0%
$PaCO_2$ change (current − first) (mmHg)	-11.0 (-18.0 to -5.0)	0.0%	-12.0 (-20.0 to -4.2)	0.0%
COPD code (retrospective)	659/2662 (24.8%)	0.0%	497/2506 (19.8%)	0.0%

Continuous variables are median (interquartile range); categorical variables are count/observed denominator (%). Missingness is before imputation. Age is capped at 90. Measurement age is the interval from the latest eligible laboratory sample to the landmark.

$PaCO_2$, arterial carbon dioxide tension; SpO_2, peripheral oxygen saturation; GCS, Glasgow Coma Scale; COPD, chronic obstructive pulmonary disease. Diagnosis codes are retrospective descriptive labels.

Characteristic	MIMIC-IV n = 2662	Missing	eICU n = 2506	Missing
Respiratory-failure code (retrospective)	1362/2662 (51.2%)	0.0%	1516/2506 (60.5%)	0.0%
Post-landmark in-hospital death	561/2662 (21.1%)	0.0%	577/2506 (23.0%)	0.0%
Continuous variables are median (interquartile range); categorical variables are count/observed denominator (%). Missingness is before imputation. Age is capped at 90. Measurement age is the interval from the latest eligible laboratory sample to the landmark.				
$PaCO_2$, arterial carbon dioxide tension; SpO_2, peripheral oxygen saturation; GCS, Glasgow Coma Scale; COPD, chronic obstructive pulmonary disease. Diagnosis codes are retrospective descriptive labels.				

Median age was 64 years in MIMIC and 65 years in eICU. Median early pH was 7.28 in both cohorts; early $PaCO_2$ was 52.0 and 55.2 mmHg, respectively. Median latest pH was 7.37 in both, with latest $PaCO_2$ of 42.0 and 45.0 mmHg. Comparable GCS was missing in 79.9% and 45.3%, and lactate in 3.8% and 50.8%.

In the broader eICU 24-hour risk set, 11322 admissions represented 192 hospitals. Mortality was 14.1% among the 11209 admissions with known outcomes. The primary paired cohort covered 22.1% of that risk set, and 38 hospitals contributed no admission with a qualifying pair and known outcome. Median paired hospital size was 8.5 admissions; 84 hospitals contributed fewer than 10. These coverage patterns and the higher mortality in the paired cohort describe the subgroup selected by repeat measurement and record availability (Additional file 1: Figure S1).

Primary model comparison and external validation

The complete MIMIC selection procedure produced outer-fold Brier scores of 0.1283 for H and 0.1278 for L (ΔBrier -0.52×10^{-3}). Per-family performances were close (Supplementary Table S5). Outer selection chose spline logistic regression in all five training sets. Final full-cohort selection chose HistGB configuration 0 for H, L and L_best: seven maximum leaf nodes, learning rate 0.03, 200 iterations, minimum leaf size 30 and L2 regularization 10.

In eICU, H and L had Brier scores of 0.1476 and 0.1478 and AUCs of 0.7545 and 0.7536. The primary paired ΔBrier was 0.27×10^{-3} (95% CI -1.10×10^{-3} to 1.69×10^{-3}). Log-loss was 0.4597 for H and 0.4609 for L. Calibration slopes were 0.880 and 0.882, and joint calibration intercepts were − 0.041 and − 0.003 (Fig. 2). Calibration-in-the-large offsets were 0.097 and 0.139. Frozen external metrics are summarized in Table 2. L_best produced the same external probabilities as L. The interval around the primary difference included both a small reduction and a small increase in prediction error.

Table 2
Frozen primary external validation

Metric	H	L	L_best
Brier score	0.1476(0.1296 to 0.1658)	0.1478(0.1297 to 0.1664)	0.1478(0.1297 to 0.1664)
AUC	0.7545 (0.7175 to 0.7884)	0.7536 (0.7135 to 0.7886)	0.7536 (0.7135 to 0.7886)
Log loss	0.4597(0.4147 to 0.5058)	0.4609(0.4148 to 0.5087)	0.4609(0.4148 to 0.5087)
Joint calibration intercept	-0.041 (-0.328 to 0.228)	-0.003 (-0.288 to 0.268)	-0.003 (-0.288 to 0.268)
Calibration slope	0.880 (0.732 to 1.028)	0.882 (0.734 to 1.031)	0.882 (0.734 to 1.031)
Calibration in the large	0.097	0.139	0.139
H adds early pH and $PaCO_2$ to the information available to L. Both use the MIMIC-selected HistGB family with separate equal-budget tuning. Independently selected L_best has identical frozen predictions to L in this run.			
Intervals condition on frozen predictions. Percentile 95% intervals use 2000 crossed hospital–patient Poisson product-weight draws for Brier/log loss and 500 for AUC/calibration. Displayed metrics are rounded; paired differences use unrounded predictions. Calibration in the large is an offset-only intercept; the joint intercept is estimated with the slope in diagnostic logistic regression.			
Primary paired difference Brier(L)−Brier(H) = 0.27×10^{-3} (95% CI -1.10×10^{-3} to 1.69×10^{-3}); positive values favor H.			

Secondary analyses

At 12 hours, development included 3624 admissions with 608 deaths. External evaluation included 3219 admissions with 655 deaths across 162 hospitals. H12/L12 Brier scores were 0.1330/0.1338, with ΔBrier 0.75×10^{-3} (95% CI -1.78×10^{-3} to 3.51×10^{-3}); AUCs were 0.7589/0.7580 and calibration slopes 0.858/0.860. Widening the repeat window yielded 4617 external admissions with 908 deaths across 168 hospitals. The frozen models produced ΔBrier 1.04×10^{-3} (95% CI -0.90×10^{-3} to 3.20×10^{-3}).

The availability-based external workflow included 10865 admissions from 10119 patients across 191 hospitals, with 1530 deaths. H24 supplied predictions for 2506 admissions (23.06%), and B_all supplied the remaining 8359. Brier scores were 0.1018 for B_all alone and 0.1020 for the workflow; Brier(B_all) −Brier(workflow) was -0.12×10^{-3} (95% CI -2.17×10^{-3} to 1.90×10^{-3}). The two strategies gave identical predictions when a repeat was unavailable, and updated predictions matched the primary H24 outputs exactly.

Secondary external comparisons are summarized in Table 3 and Fig. 3. The support audit left current support unknown in 29.1% of the MIMIC paired cohort and 91.3% of the eICU paired cohort. Adding the support indicators yielded external H_S/L_S Brier scores of 0.1503/0.1510 and ΔBrier 0.72×10^{-3} (95% CI -0.33×10^{-3} to 1.77×10^{-3}). Subgroup counts and available descriptive metrics are presented in supplementary tables. Groups with fewer than 200 admissions or fewer than 50 events/non-events were summarized by counts and events only.

Table 3
Frozen secondary external validation

Evaluation	Admissions/ deaths	Brier control/update	ΔBrier × 10^3(95% CI)	AUC control/update
12 hours [8, 12]h	3219/655 (162 hospitals)	0.1338/0.1330	0.75(-1.78 to 3.51)	0.7580/0.7589
12 hours wider window (6,12]h	4617/908 (168 hospitals)	0.1337/0.1327	1.04(-0.90 to 3.20)	0.7453/0.7502
24-hour availability workflow	10865/1530 (191 hospitals)	0.1018/0.1020	-0.12(-2.17 to 1.90)	0.7721/0.7694
Control/update are L12/H12 for the two paired evaluations and B_all/workflow for the availability analysis. The workflow uses H24 with a qualifying repeat and B_all otherwise. ΔBrier and its confidence interval are expressed in units of 10^{-3} and calculated from unrounded predictions. Positive ΔBrier favors the update.				
Secondary development conditions on the HistGB family selected in overlapping primary MIMIC data. The wider-window analysis applies the unchanged narrow-window models. Intervals use 2000 hospital–patient product-weight draws with frozen predictions; the evaluations use distinct risk sets.				

Explanations and calculator verification

Age, latest pH, lactate, GCS and creatinine had the largest mean absolute SHAP values; early pH ranked ninth and early $PaCO_2$ twelfth (Fig. 4). The importance-rank correlation between backgrounds was 0.964; additivity checks are detailed in Additional file 1 (Methods S5). Conditional joint-gas plots illustrate variation in frozen H/L predictions within the development-derived support mask (Additional file 1: Figure S3).

JavaScript browser calculations matched the Python implementation within the prespecified numerical tolerance; verification procedures and results are reported in Additional file 1 (Methods S6). The interface is publicly accessible as the hypercapnia reassessment research calculator [34]; a compatible local model file is required for calculation.

Discussion

The first blood gas provided an uncertain predictive increment beyond the latest blood gas and shared recorded clinical information at 24 hours. In 2506 external admissions, the history model H and current-state model L had Brier scores of 0.1476 and 0.1478. The paired difference, defined as Brier(L)−Brier(H), was 0.27×10^{-3} (95% CI, -1.10×10^{-3} to 1.69×10^{-3}); AUCs were 0.7545 and 0.7536. The interval includes gains and losses; equivalence was not a prespecified objective. The design combines a common prediction time, shared clinical inputs and paired external evaluation of frozen models.

The distinction between initial severity and residual prognostic information helps interpret this result. Joint pH/$PaCO_2$ abnormalities have been associated with mortality, and acidosis contributes to HACOR and NIVO [3, 6, 7]. By 24 hours, the latest blood gas and clinical measurements can already contain information about the preceding illness and its management. Earlier values can remain associated with the outcome without materially improving predictions conditional on that current information. The observed pH/$PaCO_2$ pairs describe early and later states, with acid-base changes potentially reflecting both respiratory and metabolic processes.

Related models address complementary clinical tasks. Lu and colleagues studied a diagnosis-selected acute hypercapnic respiratory failure cohort, first-day clinical burden and reported 28-day mortality [8]. Our physiologically defined cohort includes early hypercapnia with and without acidaemia and predicts death before discharge after a specified ICU landmark. Zong and colleagues used longer blood gas series and later sepsis landmarks [10]; Liu and colleagues evaluated reassessment after noninvasive ventilation initiation and a failure composite that included escalation and death [11]. Differences in risk sets and outcomes define distinct prediction tasks and prevent a direct AUC ranking.

Measurement coverage was a major determinant of the population evaluated. The external 24-hour early-hypercapnia risk set contained 11322 admissions from 192 hospitals, whereas 2506 admissions from 154 hospitals entered the known-outcome paired comparison, a coverage of 22.1%. Mortality was 23.0% in the paired group and 14.1% in the overall known-outcome risk set. These descriptive differences are consistent with measurement availability selecting a different case mix. Healthcare-process information can influence prediction in electronic records, and informative observation can distort biomarker–outcome associations [13, 35]. The availability-based workflow addressed this operational constraint: 23.06% of 10865 eligible admissions received H, and the remainder retained B_all. Its external Brier difference was -0.12×10^{-3} (95% CI, -2.17×10^{-3} to 1.90×10^{-3}). This estimates performance under observed sampling practice; estimating the benefit of additional sampling requires a design that changes or standardises that practice.

The 12-hour analyses extended the question to earlier reassessment. In the separately defined 12-hour risk set, the external paired Brier difference was 0.75×10^{-3} (95% CI, -1.78×10^{-3} to 3.51×10^{-3}). Applying the frozen 12-hour models to the wider (6,12]-hour repeat window increased external coverage to 4617 admissions and yielded a difference of 1.04×10^{-3} (95% CI, -0.90×10^{-3} to 3.20×10^{-3}). This wider-window analysis assesses transport to a broader measurement window. The 12- and 24-hour

populations overlap and condition on different survival and ICU-stay requirements, so their performance estimates answer separate reassessment questions and do not identify an optimal sampling time.

External calibration remained relevant despite similar H and L discrimination. The 24-hour calibration slopes were 0.880 and 0.882; the 12-hour slopes were 0.858 and 0.860. Calibration regression was diagnostic and left the frozen predictions unchanged. Evaluation guidance recommends examining absolute-risk agreement across the prediction range: intercept and slope summaries complement assessment throughout that range [14, 31]. Differences in case mix and data collection also affect how performance in a new population should be interpreted [36]. A nearby COPD ICU mortality study provided a web calculator yet reported a negative incremental net benefit relative to SAPS II in eICU [37]. This provides a practical reason to assess transport and decision consequences separately from model presentation.

SHAP showed how the frozen H pipeline allocated its logit-risk predictions [32]. Age, latest pH, lactate, GCS and creatinine had the largest average attributions; first pH and $PaCO_2$ ranked ninth and twelfth. Background and prespecified blood gas grouping analyses supported consistency of these calculations; additivity diagnostics are reported in Additional file 1 (Methods S5). With dependent predictors, independent masking can generate unrealistic combinations and change Shapley attributions, making the background and grouping choices relevant to their interpretation [38]. Conditional joint blood gas plots complemented the attributions within a development-data support mask and a fixed reference clinical profile; formal interaction and treatment-effect estimates require separate analyses.

The browser calculator provides a research interface for inspecting frozen paired predictions. Numerical verification is detailed in Additional file 1 (Methods S6). Model-presentation guidance recommends verifying that a calculator reproduces its underlying model and clearly defines its intended population [39]. The public interface contains generic inference code; users load a model package locally, and inputs and model parameters remain in browser memory. Researchers can contact the corresponding author to request inspection of the fitted model package, subject to a PhysioNet-controlled access arrangement under the source-data agreement [33]. Clinical use additionally requires task-based usability evaluation and assessment of decision consequences in the intended setting. A specified clinical action and the relative consequences of false-positive and false-negative predictions would provide the basis for evaluating net benefit across relevant risk thresholds [40].

The physiological landmark population includes hypercapnia across causes of illness and respiratory support. The pooled comparison evaluates the added predictive information of early gases at a common reassessment time with the same recorded current information. Its value within individual etiologies or support modes remains unresolved. Ventilatory management can affect both blood gas changes and prognosis. Cross-database support mapping was incomplete, with unknown states in 29.1% of MIMIC and 91.3% of eICU admissions. In the support sensitivity, adding the same four recorded support categories to H and L yielded an external ΔBrier of 0.72×10^{-3} (95% CI, -0.33×10^{-3} to 1.77×10^{-3}; Supplementary Table S9), whose interval also spanned zero.

The retrospective data impose several limits. Discharge status and offsets in eICU provide proxies for subsequent death timing, and revised laboratory timestamps cannot reconstruct initial result release. Same-offset gas pairing lacks a specimen identifier. GCS and lactate were retained as prespecified clinical inputs describing neurological and metabolic status relevant to prognosis. GCS was missing in 79.9% of MIMIC and 45.3% of eICU paired admissions, and eICU lactate in 50.8%. These rates reflect measurement and record availability together with timing, quality and comparability requirements. Training-fold median imputation retained these admissions, while missingness indicators allowed the models to use the absence of an eligible recorded value. Such patterns can contain information about measurement and documentation practice; differences between hospitals remain a source of transport uncertainty [41]. Algorithm comparison used a finite configuration budget; secondary development was conditional on the family selected in overlapping MIMIC data. Hospital–patient product-weight resampling reflects clustered sampling variation for frozen predictions, following methods for crossed data arrays [42], and excludes uncertainty from repeating model development. Prospective reassessment with standardised measurements and independently validated outcomes would address these data and transport limitations.

Conclusions

This study places early blood gas history in a direct comparison with contemporary reassessment. Across the primary and prespecified external analyses, its added probability accuracy remained uncertain. Clinical evaluation should now prioritise the completeness and timing of current information, calibration in the intended population and the decision for which a risk estimate will be used.

Abbreviations

ABG: arterial blood gas; AUC: area under the receiver-operating-characteristic curve; CI: confidence interval; COPD: chronic obstructive pulmonary disease; eICU-CRD: eICU Collaborative Research Database; GCS: Glasgow Coma Scale; H: history model; HistGB: histogram gradient boosting; ICU: intensive care unit; L: current-information model; L_best: independently selected current-information model; MAP: mean arterial pressure; MIMIC-IV: Medical Information Mart for Intensive Care IV; $PaCO_2$: arterial carbon dioxide tension; SHAP: Shapley Additive Explanations; SpO_2: peripheral oxygen saturation; SVM: support vector machine.

Declarations

Ethics approval and consent to participate

This study adhered to the ethical principles of the Declaration of Helsinki. The analyses were performed using preexisting, de-identified data from two publicly accessible databases: MIMIC-IV (v3.1) and eICU-CRD (v2.0), both hosted by PhysioNet. The establishment of the MIMIC-IV database was approved by the Institutional Review Boards of the Massachusetts Institute of Technology and Beth Israel Deaconess

Medical Center, with a waiver of informed consent granted due to data anonymization [20]. The eICU-CRD resource was certified as meeting HIPAA Safe Harbor de-identification standards by Privacert, exempting it from institutional review board oversight [21]. The authors completed the required human subjects research training (CITI Program) and signed the data-use agreements to access the source records [20, 22]. Because this study exclusively involved secondary analysis of fully de-identified publicly available data with no individual intervention or contact, additional ethical review and informed consent by the local institutional ethics committee were not required in accordance with institutional policies and national guidelines.

Clinical trial number

Not applicable.

Consent for publication

Not applicable. No identifiable individual information is presented.

Availability of data and materials

MIMIC-IV 3.1 and eICU 2.0 are available through PhysioNet subject to credentialing and their data-use agreements [20, 22]. Extraction and modeling code, operational definitions and aggregate figure and table data accompany this manuscript in Additional file 2. Re-execution requires authorized source access and the documented preparation steps. The research interface is publicly accessible as the hypercapnia reassessment research calculator [34]. A compatible locally loaded model file is required for calculation. Frozen model configuration files are included in Additional file 2. Requests for research inspection of fitted parameters should be directed to the corresponding author; provision requires a PhysioNet-controlled access arrangement under the source-data agreement [33].

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the Postgraduate Quality Improvement and Innovation Project of Shandong University of Traditional Chinese Medicine (Top-notch Innovation Program; Graduate Training Fund of the First Clinical Medical College, Grant No. 230043). The funding body had no role in study design, data collection, analysis, data interpretation or manuscript preparation.

Authors' contributions

JX and WQ conceived and designed the study. JX and XL contributed to data curation, extraction and code implementation. JX, PL and XC contributed to statistical analysis and computational verification. JX drafted the manuscript and prepared figures and tables. WQ provided supervision, resources, project administration and critical revision for clinical and intellectual content. All authors read and approved the final manuscript.

Acknowledgements

We thank the teams who created and maintain MIMIC-IV, eICU and PhysioNet, and the participating hospitals that contributed the source records.

References

1. Rochwerg B, Brochard L, Elliott MW, Hess D, Hill NS, Nava S, et al. Official ERS/ATS clinical practice guidelines: noninvasive ventilation for acute respiratory failure. Eur Respir J. 2017;50(2):1602426. 10.1183/13993003.02426-2016.
2. Davidson AC, Banham S, Elliott M, Kennedy D, Gelder C, Glossop A, et al. BTS/ICS guideline for the ventilatory management of acute hypercapnic respiratory failure in adults. Thorax. 2016;71(Suppl 2):ii1–35. 10.1136/thoraxjnl-2015-208209.
3. Tiruvoipati R, Pilcher D, Buscher H, Botha J, Bailey M. Effects of Hypercapnia and Hypercapnic Acidosis on Hospital Mortality in Mechanically Ventilated Patients. Crit Care Med. 2017;45(7):e649–56. 10.1097/CCM.0000000000002332.
4. Nin N, Muriel A, Peñuelas O, Brochard L, Lorente JA, Ferguson ND, et al. Severe hypercapnia and outcome of mechanically ventilated patients with moderate or severe acute respiratory distress syndrome. Intensive Care Med. 2017;43(2):200–8. 10.1007/s00134-016-4611-1.
5. Maamar A, Delamaire F, Reizine F, Lesouhaitier M, Painvin B, Quelven Q, et al. Impact of Arterial CO2 Retention in Patients With Moderate or Severe ARDS. Respir Care. 2023;68(5):582–91. 10.4187/respcare.10507.
6. Duan J, Wang S, Liu P, Han X, Tian Y, Gao F, et al. Early prediction of noninvasive ventilation failure in COPD patients: derivation, internal validation, and external validation of a simple risk score. Ann Intensive Care. 2019;9(1):108. 10.1186/s13613-019-0585-9.
7. Hartley T, Lane ND, Steer J, Elliott MW, Sovani MP, Curtis HJ, et al. The Noninvasive Ventilation Outcomes (NIVO) score: prediction of in-hospital mortality in exacerbations of COPD requiring assisted ventilation. Eur Respir J. 2021;58(2):2004042. 10.1183/13993003.04042-2020.
8. Lu C, Lin J, Yue Y, Fu J, Zhang G. Interpretable machine learning based on the Charlson comorbidity index predicts 28-day mortality in acute hypercapnic respiratory failure. Sci Rep. 2026;16(1):3335. 10.1038/s41598-025-33251-9.

9. van Houwelingen HC. Dynamic Prediction by Landmarking in Event History Analysis. Scand J Stat. 2007;34(1):70–85. 10.1111/j.1467-9469.2006.00529.x.
10. Zong LR, Bi QY, Liu Y, Lu SJ, Li CY, Wei ZK, et al. Dynamic prognostic prediction in sepsis using longitudinal blood gas trajectories: development and external validation. Front Med (Lausanne). 2026;13:1852841. 10.3389/fmed.2026.1852841.
11. Liu T, Yan X, Lu J. Two-hour dynamic reassessment of noninvasive ventilation failure risk in acute exacerbation of chronic obstructive pulmonary disease with hypercapnic respiratory failure: development and temporal validation. Front Med (Lausanne). 2026;13:1920677. 10.3389/fmed.2026.1920677.
12. Pencina MJ, D'Agostino RB, Sr, Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157–72. 10.1002/sim.2929.
13. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479. 10.1136/bmj.k1479.
14. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. 10.1136/bmj-2023-074820.
15. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
16. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. 10.1136/bmj-2024-082505.
17. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. J Am Stat Assoc. 2007;102(477):359–78. 10.1198/016214506000001437.
18. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–38. 10.1097/EDE.0b013e3181c30fb2.
19. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1. 10.1038/s41597-022-01899-x.
20. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S et al. MIMIC-IV (version 3.1). PhysioNet; 2024. https://doi.org/10.13026/kpb9-mt58.
21. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. 10.1038/sdata.2018.178.
22. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). PhysioNet; 2019. https://doi.org/10.13026/C2WM1R

23. Pollard T, Moody BE, Lehman LWH, Gow BJ, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. Nat Health. 2026;1(8):792–5. 10.1038/s44360-026-00096-z.
24. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32. 10.1023/A:1010933404324.
25. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189–232. 10.1214/aos/1013203451.
26. Chen T, Guestrin C, XGBoost:. A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: Association for Computing Machinery; 2016. pp. 785–794. 10.1145/2939672.2939785
27. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97. 10.1007/BF00994018.
28. Platt JC. Probabilities for SV Machines. In: Smola AJ, Bartlett P, Schölkopf B, Schuurmans D, editors. Advances in Large-Margin Classifiers. Cambridge, MA: MIT Press; 2000. pp. 61–74. 10.7551/mitpress/1113.003.0008.
29. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. New York, NY: Association for Computing Machinery; 2005. pp. 625–632. 10.1145/1102351.1102430
30. Cawley GC, Talbot NLC. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J Mach Learn Res. 2010;11:2079–107.
31. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76. 10.1016/j.jclinepi.2015.12.005.
32. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst. 2017;30:4765–74.
33. PhysioNet. Guidelines for creating datasets and models from MIMIC. 2024. https://physionet.org/news/post/mimic-derived-datasets-models/ Accessed 9 Oct 2026.
34. Xu J, Liang X, Li P, Cai X, Qian W. Hypercapnia reassessment research calculator. 2026. https://xujitao2002.github.io/hypercapnia-calculator/ Accessed 9 Oct 2026.
35. Harton J, Mitra N, Hubbard RA. Informative presence bias in analyses of electronic health records-derived data: a cautionary note. J Am Med Inf Assoc. 2022;29(7):1191–9. 10.1093/jamia/ocac050.
36. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279–89. 10.1016/j.jclinepi.2014.06.018.
37. Li Y, Bu X, Geng Z, Dai N, Fu B, Deng R, et al. Interpretable machine learning for predicting in-hospital mortality in COPD ICU patients: a rigorous validation across time and geography. Respir Res. 2026;27(1):283. 10.1186/s12931-026-03702-6.
38. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell. 2021;298:103502.

10.1016/j.artint.2021.103502.

39. Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. BMJ. 2019;365:l737. 10.1136/bmj.l737.
40. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Mak. 2006;26(6):565–74. 10.1177/0272989X06295361.
41. Sisk R, Sperrin M, Peek N, van Smeden M, Martin GP. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. Stat Methods Med Res. 2023;32(8):1461–77. 10.1177/09622802231165001.
42. Owen AB, Eckles D. Bootstrapping data arrays of arbitrary order. Ann Appl Stat. 2012;6(3):895–927. 10.1214/12-AOAS547.

Figures

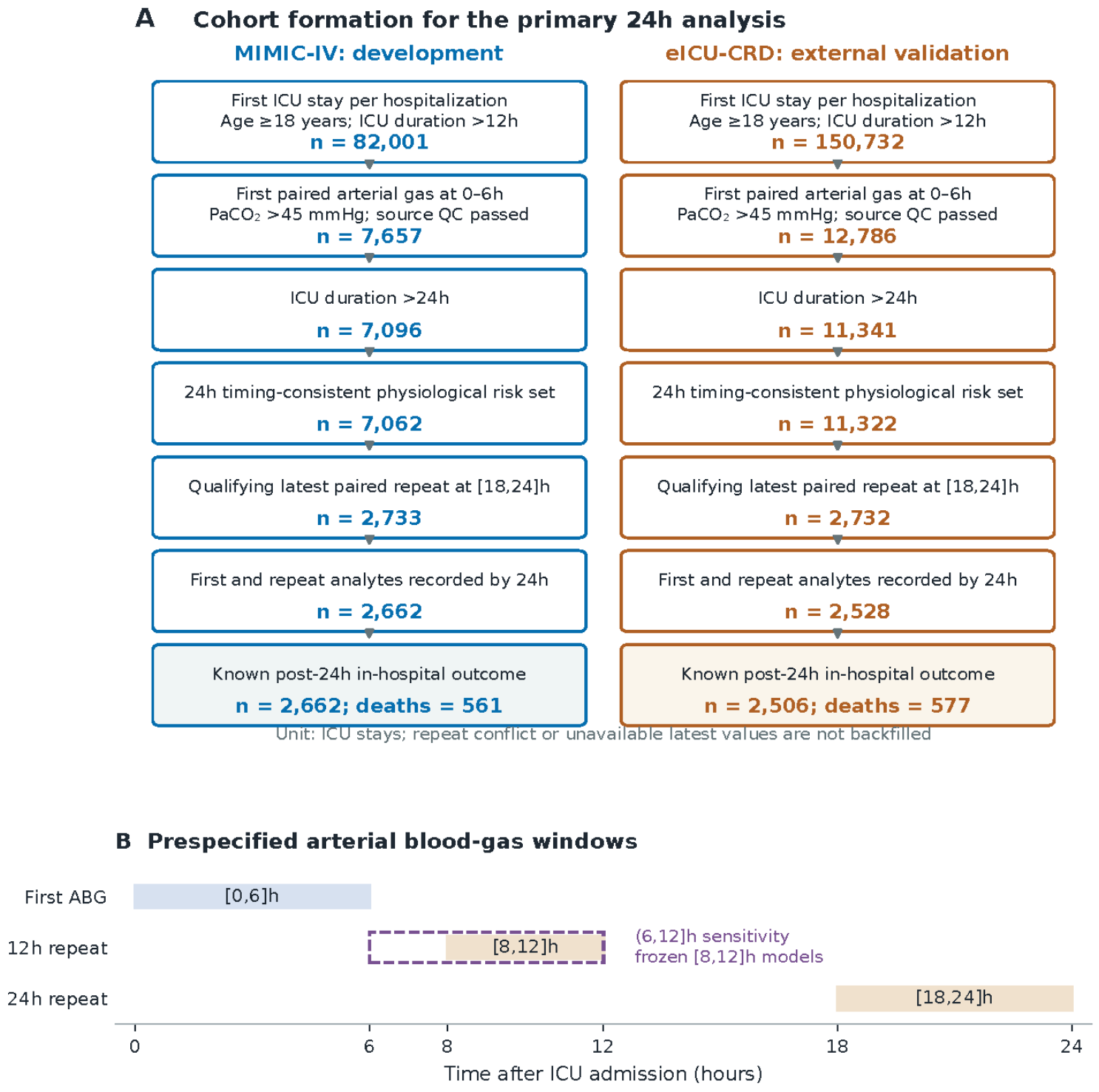

Figure 1

Cohort selection and prespecified blood gas landmarks

A, parallel formation of the development and external 24-hour cohorts. Counts are ICU admissions. The initial frame requires the first ICU stay per hospitalization, adult age and ICU duration strictly greater than 12 hours; it is not the complete database denominator. The earliest eligible paired arterial gas in [0,6] hours must have $PaCO_2$ >45 mmHg. The primary analysis additionally requires ICU duration greater than 24 hours, timing consistency, an eligible latest repeat in [18,24] hours, recorded first and repeat analytes

by the node, and a known post-node hospital outcome. Final cohorts contain 2662/2506 stays and 561/577 deaths in MIMIC/eICU. Intermediate gates retain their source-QC definitions. B, prespecified first, 12-hour and 24-hour gas windows. The dashed outline denotes the (6,12] hour sensitivity using unchanged [8,12] hour models. Conflicting or unavailable latest repeats are not replaced by older measurements. Recorded/revision timestamps are retrospective availability proxies. Eligibility defines a physiological hypercapnia population, including patients with and without respiratory failure or COPD diagnoses. ABG, arterial blood gas.

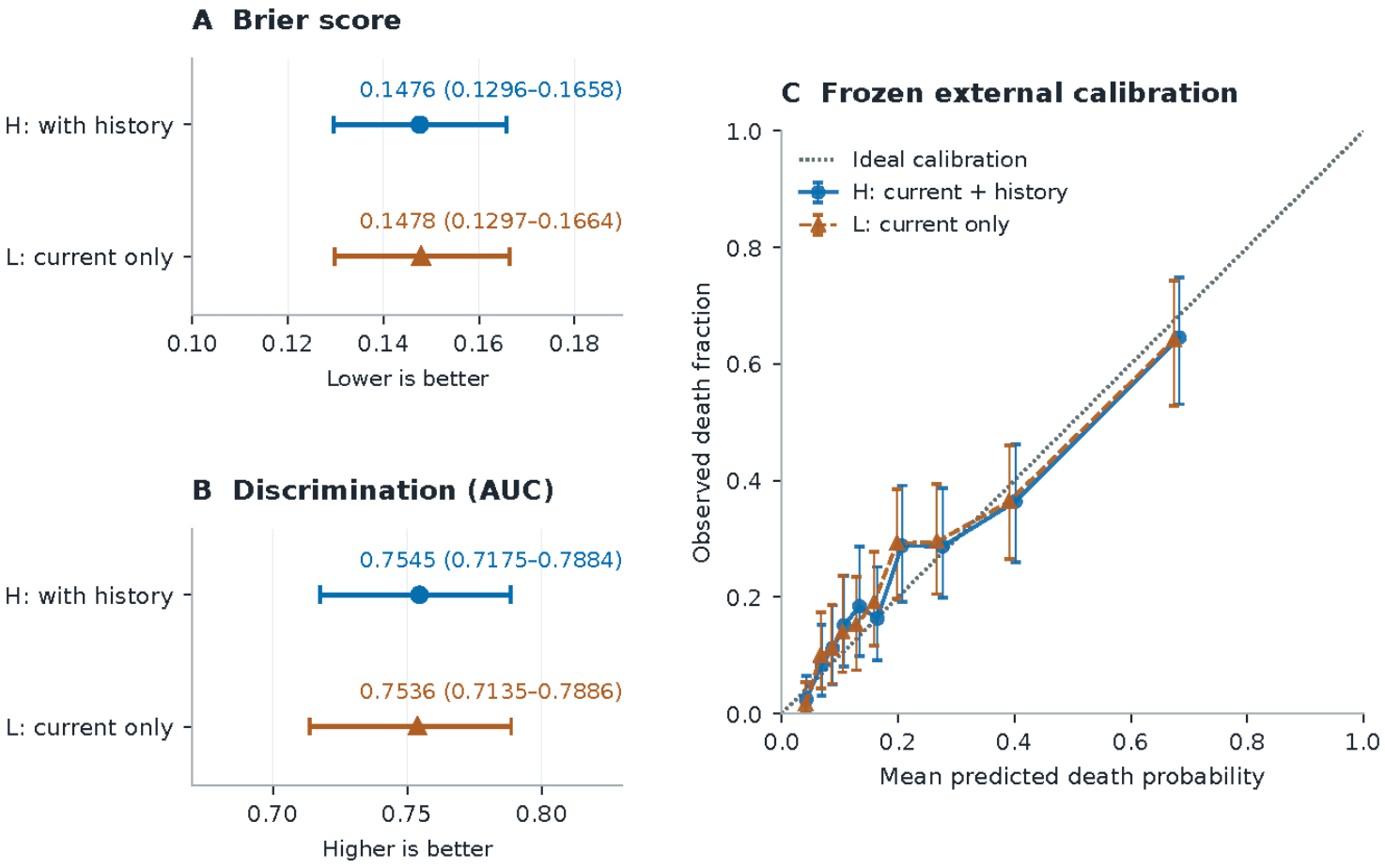

Figure 2

Frozen primary model performance and calibration in eICU

The same 2506 eICU admissions, with 577 deaths across 154 hospitals, supplied both evaluations. H adds first pH/$PaCO_2$ to the clinical inputs and current gas supplied to L. Both use the MIMIC-selected HistGB family with separate equal-budget tuning and development-only sigmoid calibration. A, Brier score (lower is better). B, area under the receiver-operating-characteristic curve (AUC; higher is better). Points and horizontal 95% percentile confidence intervals use existing crossed hospital–patient independent Poisson product-weight draws: 2000 for Brier and 500 for AUC. Figure 3 shows the paired increment. C, original model-specific prediction-quantile bins show mean predicted death probability and observed death fraction, with 95% percentile observed-fraction intervals from 500 draws and bin membership fixed. Lines connect the saved bin summaries. Both axes span 0–1 with equal physical

scale; the diagonal denotes ideal calibration. Probabilities were not recalibrated in eICU. Intervals condition on frozen models and exclude redevelopment uncertainty. The independently selected L_best has identical predictions to L.

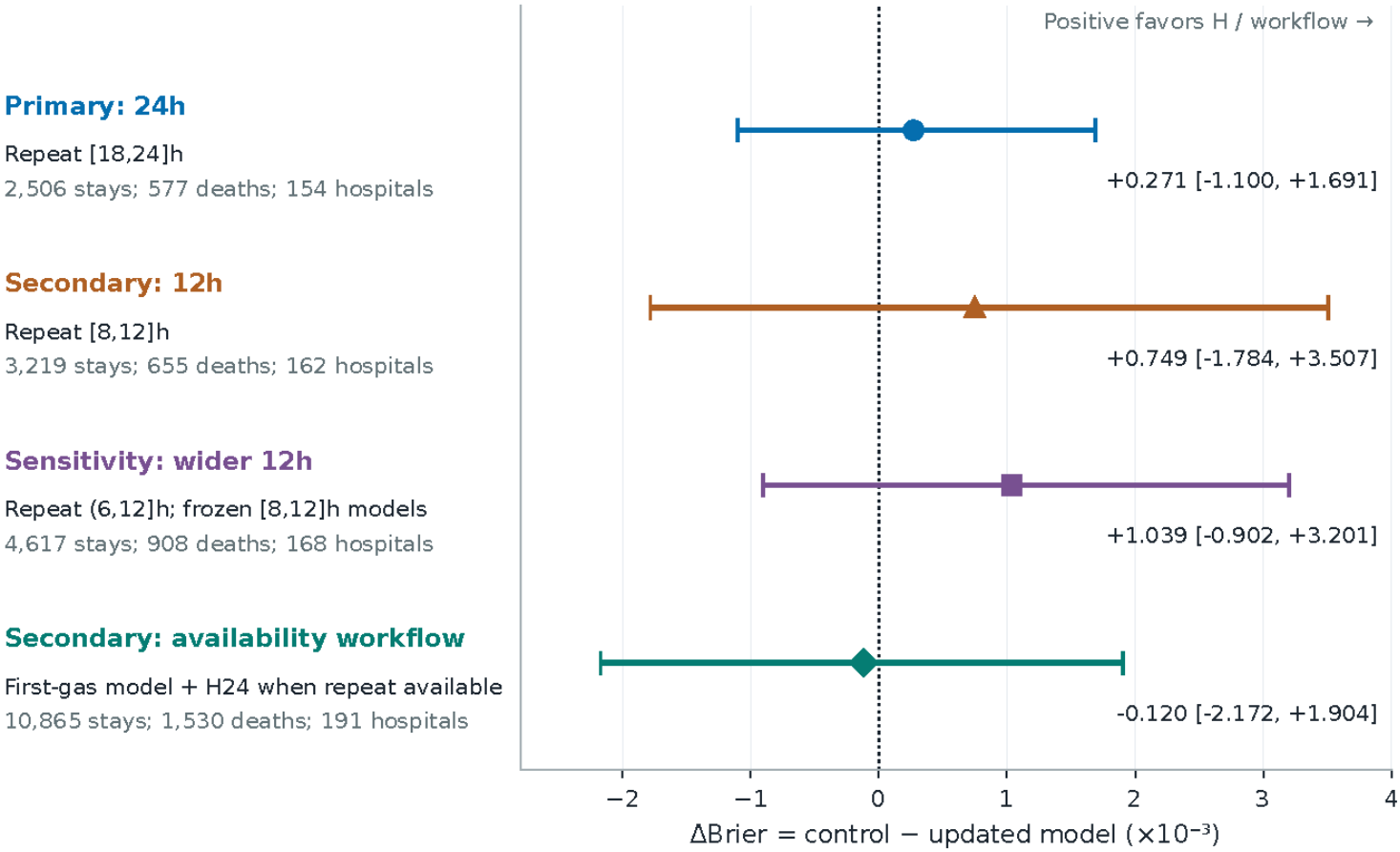

Figure 3

Paired incremental Brier performance across external evaluations

Points and horizontal lines show existing paired differences and 95% percentile confidence intervals from 2000 crossed hospital–patient Poisson product-weight draws with frozen predictions. ΔBrier is control minus updated-model Brier score; positive values favor H or the availability workflow. Displayed values are multiplied by 10^3, so one axis unit is 0.001 Brier score. Control/update are L/H at 24 hours, L12/H12 at 12 hours, and B_all/workflow for the 24-hour availability evaluation. B_all uses clinical inputs and the first gas; the workflow uses H24 when a qualifying repeat exists and B_all otherwise. The wider (6,12] hour sensitivity retains the models developed using [8,12] hour repeats. Sample sizes, event counts and hospitals are specific to each row. All intervals cross zero. Each estimate applies to its own, partly overlapping landmark or availability risk set. Secondary family choice conditions on the primary MIMIC-selected family. External and wider-window analyses use the unchanged development models.

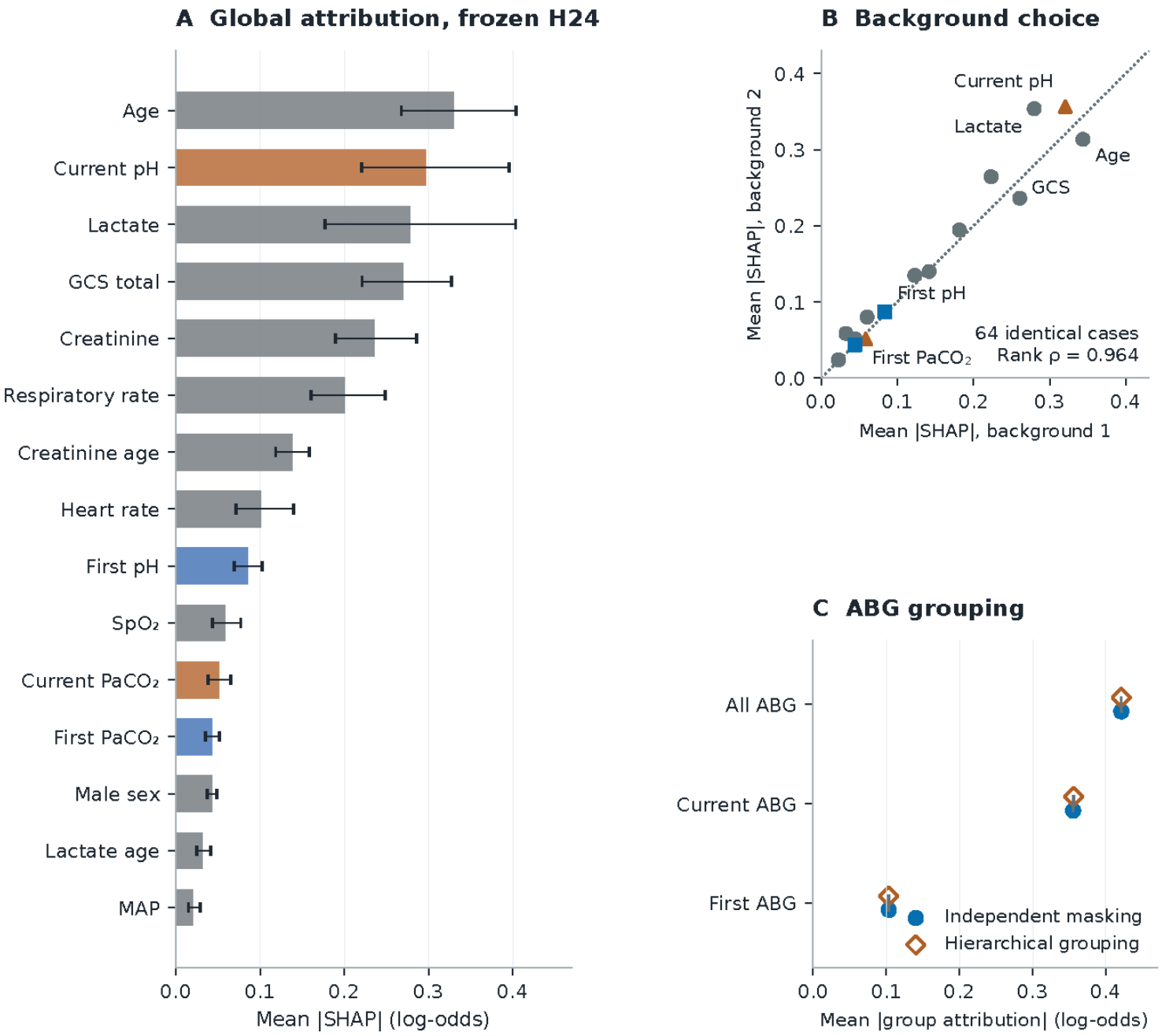

Figure 4

SHAP attributions and explanation sensitivity for the frozen H24 pipeline

SHAP explains the logit of the complete calibrated H24 probability, clipped at 10^{-12} and $1-10^{-12}$, including preprocessing and missingness handling. A, mean absolute attribution in 128 randomly selected eICU admissions using 32 MIMIC background admissions. Horizontal 95% percentile intervals use 1000 crossed hospital–patient explanation-sample weight draws with the model, background and selected cases fixed. Blue/orange bars identify first/current gas; gray bars identify other clinical inputs. These intervals describe sampling variation in explanations of the frozen model. B, mean absolute importance under two disjoint 32-admission backgrounds on the same 64 external cases; the diagonal denotes equality and importance-rank Spearman correlation is 0.964. C, mean absolute group attribution

in 32 cases comparing independent-mask SHAP with hierarchical Owen values. Small vertical marker offsets expose near-coincident means and have no quantitative meaning. First/current/all-gas coalitions overlap; their absolute magnitudes are not additive disjoint partitions. These diagnostics characterize the frozen prediction function and explanation sensitivity. ABG, arterial blood gas; GCS, Glasgow Coma Scale; MAP, mean arterial pressure.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1SupplementaryMethods.docx
- AdditionalFile2CodeandAggregates01.zip